

Давиденко Анатолій Миколайович

доктор технічних наук, кафедра комп'ютерної інженерії

Державний університет інформаційно-комунікаційних технологій, Київ, Україна

ORCID ID: 0000-0001-6466-1690

davidenkoan@gmail.com

Азарний Дмитро Ігорович

аспірант факультету радіофізики, електроніки та комп'ютерних систем

Київський національний університет імені Тараса Шевченка, Київ, Україна

ORCID ID: 0009-0000-9549-8873

dazarny@gmail.com

Висоцька Олена Олександрівна

кандидат технічних наук, кафедра кібербезпеки

Національний університет «Київський авіаційний інститут», Київ, Україна

ORCID ID: 0000-0002-9543-1385

Lek_vys@ukr.net

ЧАСОВЕ ПОКРИТТЯ ВІДЕО У ВИЯВЛЕННІ ДИПФЕЙКІВ: ТОЧНІСТЬ ТА ОБЧИСЛЮВАЛЬНІ ВИТРАТИ

Анотація. Детектори відеодипфейків часто аналізують не весь запис, а один або кілька коротких фрагментів. Якщо ознаки підробки розподілені в часі нерівномірно, результат для одного фрагмента може не характеризувати відео в цілому. Водночас збільшення кількості фрагментів майже пропорційно підвищує обчислювальні витрати. Метою роботи було визначити, скільки рівномірно розміщених фрагментів потрібно для надійного оцінювання коротких відео та після якої кількості подальше збільшення часового покриття перестав давати практично помітний приріст. Дослідження виконано з незмінним детектором, що поєднує просторові та частотні ознаки кадрів із трансформерним аналізом їх послідовності. Три повторно навчені моделі об'єднано усередненням і перевірено на окремій зовнішній вибірці з 200 узгоджених пар справжніх і підроблених відео. Для кожного відео один раз оброблено дев'ять рівномірно розміщених 16-кадрових фрагментів. Режим з одним, трьома та п'ятьма фрагментами сформовано як вкладені підмножини тієї самої дев'ятичовкової сітки. Завдяки цьому відмінності не залежали від випадкового вибору кадрів або повторного запуску моделі. ROC AUC становив 0,8349 для одного центрального фрагмента, 0,8688 для трьох, 0,8724 для п'яти та 0,8732 для дев'яти. Дев'ять фрагментів підвищили AUC відносно одного на 0,0383; р-значення дорівнювало 0,001. Три фрагменти забезпечили 88,5% цього приросту за третини максимального обчислювального бюджету, а п'ять - 97,7% за п'яти дев'ятих бюджету. Підтвердженого поліпшення між п'ятьма і дев'ятьма фрагментами не виявлено. Мінливість оцінок між фрагментами у підроблених відео була приблизно у 1,9 раза вищою, ніж у справжніх. Медіана, максимум і середнє двох найбільших оцінок не перевершили звичайного середнього. Один центральний фрагмент недостатньо представляє коротке відео за умов перенесення між датасетами. Три фрагменти є доцільним режимом за жорсткого обмеження ресурсів, а п'ять - обережним вибором поблизу виявленого плато. Висновок не поширюється автоматично на багатохвилинні записи, для яких потрібне окреме дослідження адаптивного часового відбору.

Ключові слова: цифрова форензика, часове покриття, ROC AUC, міждатасетне оцінювання, часова неоднорідність, агрегація прогнозів, аналіз відео.

Davydenko Anatolii

Doctor of Technical Sciences, Department of Computer Engineering

State University of Information and Communication Technologies, Kyiv, Ukraine

ORCID ID: 0000-0001-6466-1690

davidenkoan@gmail.com

Azarnyi Dmytro

PhD student, Faculty of Radiophysics, Electronics and Computer Systems

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

ORCID ID: 0009-0000-9549-8873

dazarny@gmail.com

Vysotska Olena

Candidate of Technical Sciences, Department of Cybersecurity

National University 'Kyiv Aviation Institute', Kyiv, Ukraine

ORCID ID: 0000-0002-9543-1385

Lek_vys@ukr.net

© 2026 Давиденко А. М., Азарний Д.І., Висоцька О.О.. Цей матеріал ліцензовано за умовами CC BY 4.0.

<https://creativecommons.org/licenses/by/4.0/>

VIDEO TEMPORAL COVERAGE IN DEEFAKE DETECTION: ACCURACY AND COMPUTATIONAL COST

Abstract. Video deepfake detectors commonly analyse one or several short clips instead of the entire recording. When forgery evidence is unevenly distributed over time, one clip may not represent the video as a whole, while processing more clips increases computational cost almost proportionally. This study determined how many uniformly placed clips are needed for reliable evaluation of short videos and where additional temporal coverage ceases to provide a practically meaningful gain. The detector was fixed throughout the experiment and combined spatial and frequency-domain frame features with Transformer-based sequence analysis. Predictions from three independently trained models were averaged and evaluated on a separate external sample of 200 matched real-fake video pairs. Nine uniformly spaced 16-frame clips were processed once for every video. The one-, three-, and five-clip modes were reconstructed as nested subsets of the same nine-position grid, so observed differences were not caused by resampling or another model run. ROC AUC was 0.8349 for one central clip, 0.8688 for three clips, 0.8724 for five clips, and 0.8732 for nine clips. Nine clips increased AUC by 0.0383 over one clip, with $p = 0.001$. Three clips retained 88.5% of this gain at one third of the maximum computational budget; five clips retained 97.7% at five ninths of that budget. No confirmed improvement was observed between five and nine clips. Prediction variability across clips was approximately 1.9 times higher for fake videos than for real videos. The median, maximum, and mean of the two largest clip scores did not outperform the arithmetic mean. A single central clip therefore under-represents short videos under cross-dataset evaluation. Three clips are suitable under a strict resource limit, whereas five clips are a conservative choice near the observed plateau. The finding should not be transferred directly to multi-minute recordings, which require a separate evaluation of duration-aware and adaptive temporal sampling.

Keywords: digital forensics, temporal coverage, ROC AUC, cross-dataset evaluation, temporal heterogeneity, prediction aggregation, inference efficiency.

1. Вступ

Сучасні засоби синтезу та заміни обличчя дають змогу створювати правдоподібні відео, які можуть бути використані для дезінформації, шахрайства або підміни особи. Автоматичне виявлення таких матеріалів розвивається як окремий напрям цифрової форензики, однак висока точність усередині одного еталонного набору ще не гарантує надійної роботи на відео з іншого джерела [1]. Результат залежить не лише від архітектури нейромережі, а й від підготовки вхідних даних та правил, за якими оцінки окремих кадрів або фрагментів перетворюють на рішення для цілого відео. Уніфіковані порівняння DeepfakeBench показали, що відмінності у конвеєрах підготовки та оцінювання можуть заважати коректному зіставленню детекторів [2].

Відео містить значно більше інформації, ніж один кадр або один короткий фрагмент. Положення голови, міміка, освітлення, розмиття руху, перекриття обличчя та якість кодування змінюються в часі. Ознаки підробки також можуть проявлятися нерівномірно. У частині кадрів межа заміни або неприродна динаміка помітні, тоді як в інших кадрах обличчя майже нерухоме, мале або частково закрите. Якщо детектор аналізує лише середину відео, його оцінка може залежати від випадково вибраної події, а не від властивостей запису в цілому.

2. Постановка проблеми

Повне покадрове оцінювання усуває ризик пропустити інформативну ділянку, але для глибоких моделей воно часто є надмірно дорогим. Кожен додатковий фрагмент потребує декодування, пошуку обличчя, формування входу та прямого проходження нейромережі. Тому практична задача полягає не у використанні максимально можливої кількості кадрів, а у визначенні найменшого часового бюджету, після якого подальші витрати дають лише незначний приріст.

Попередні роботи довели корисність часових ознак для виявлення маніпуляцій [3-9], проте здебільшого порівнювали різні архітектури або способи навчання. Значно менше уваги приділено контрольованому питанню: як змінюється оцінка тієї самої вже навченої моделі, якщо рішення для відео будувати за одним, трьома, п'ятьма або дев'ятьма рівномірно розміщеними фрагментами. Без такого контролю приріст неможливо однозначно пов'язати з часовим покриттям, оскільки одночасно можуть змінюватися мережа, аугментації, кадрування та навчальні дані.

3. Аналіз останніх досліджень і публікацій

3.1. Часові ознаки у виявленні підроблених відео

Перші відеодетектори поєднували згортокве виділення ознак кадру з рекурентною мережею, яка накопичувала інформацію вздовж послідовності. В одному з перших таких підходів кадри кодували згортковою мережею, а їх послідовність аналізували рекурентним блоком [3]. Подальше порівняння кількох рекурентно-згорткових стратегій показало перевагу аналізу послідовності над незалежним аналізом окремих кадрів [4]. Ці підходи сформували базову схему: локальні просторові ознаки спочатку виділяють для кожного кадру, після чого об'єднують у часі.

Подальші методи намагалися перейти від низькорівневих артефактів до поведінкових або довготривалих закономірностей. LipForensics аналізує рухи рота за допомогою просторово-часової мережі, попередньо навченої на читанні з губ [5]. FTCN обмежує просторову складову згорток, щоб мережа не ігнорувала часову неузгодженість на користь легших просторових ознак, а довші залежності обробляє трансформер [6]. RealForensics використовує самоконтрольоване навчання на справжніх відео мовлення, щоб отримати щільні у часі подання природної міміки та руху [7].

AltFreezing розглядає просторові й часові ознаки як взаємодоповнювальні та по черзі заморожує відповідні групи ваг під час навчання [8]. TALL перетворює кілька кадрів фрагмента на компактну мозаїку, зберігаючи частину просторово-часових зв'язків за меншої вартості, ніж у багатьох тривимірних мережах [9]. У

сукупності ці результати підтверджують, що часова інформація корисна. Однак вони не дають прямої відповіді, скільки віддалених одна від одної ділянок відео потрібно переглянути вже навченому детектору.

3.2. Вибір фрагментів і рішення для цілого відео

У задачах розпізнавання дій проблему довгого відео розв'язували розрідженим рівномірним відбором. Temporal Segment Networks поділяє запис на часові частини, бере представника з кожної та об'єднує їх на рівні відео [10]. Такий принцип забезпечує ширше покриття без повного покадрового аналізу. Для виявлення дипфейків він є привабливим з тієї самої причини: кілька віддалених фрагментів зменшують залежність від однієї локальної сцени.

У літературі немає єдиного правила формування відеооцінки. Застосовують середнє, максимум, голосування, рекурентне об'єднання, увагу або навчені ваги кадрів. Максимум підвищує вплив найпідозрілішої ділянки, але водночас робить результат чутливим до одиничного хибного спрацювання. Медіана стійкіша до викидів, проте може пригнічувати коротку інформативну подію. Просте середнє не має додаткових параметрів і використовує всі спостереження, тому є природною базовою стратегією.

Порівняння різної кількості фрагментів має дві методичні пастки. По-перше, незалежний випадковий відбір для кожного режиму додає шум: різниця може виникнути через вдалі або невдалі часові позиції. По-друге, вибір кількості фрагментів за тією самою тестовою вибіркою завищує оцінку. У цій роботі першу проблему усунуто вкладеними фіксованими підмножинами єдиної дев'ятиточкової сітки, а другу - фіксуванням основного порівняння до відкриття фінального тесту.

Невирішеним залишається практичне питання про точку насичення. Якщо три фрагменти дають майже той самий результат, що дев'ять, додаткові проходи моделі можуть бути невиправданими. Якщо ж оцінки сильно змінюються вздовж підробленого відео, один центральний фрагмент є ненадійним навіть для коротких записів. Саме спільний аналіз ROC AUC, внутрішньовідеової мінливості та відносної вартості становить предмет цієї роботи.

4. Мета і задачі дослідження

Метою роботи є кількісно оцінити вплив кількості рівномірно розміщених відеофрагментів на міждасетний ROC AUC незмінного детектора та визначити співвідношення між приростом і обчислювальним бюджетом. Для досягнення мети поставлено чотири завдання: порівняти вкладені режими з одним, трьома, п'ятьма і дев'ятьма фрагментами; перевірити, чи зберігається результат за різної якості відео; оцінити мінливість прогнозів уздовж відео; зіставити просте усереднення з альтернативними правилами об'єднання.

Особливістю експерименту є те, що модель, її ваги, спосіб кадрування та фінальна тестова вибірка залишалися незмінними. Усі чотири режими відновлено з одного набору дев'яти прогнозів для кожного відео. Отже, отримані відмінності відображають саме часовий бюджет. Додатково показано, яка частка максимального спостережуваного приросту досягається трьома і п'ятьма фрагментами та як цей ефект пов'язаний з неоднорідністю оцінок у часі.

5. Результати дослідження

5.1. Дизайн і протокол дослідження

Дослідження мало два послідовні етапи. На першому етапі навчали детектор і визначали всі його налаштування без використання фінальної тестової вибірки. На другому етапі незмінну модель один раз застосували до окремої зовнішньої вибірки. Для кожного відео отримали дев'ять прогнозів у фіксованих часових позиціях, після чого з них сформували чотири режими оцінювання.

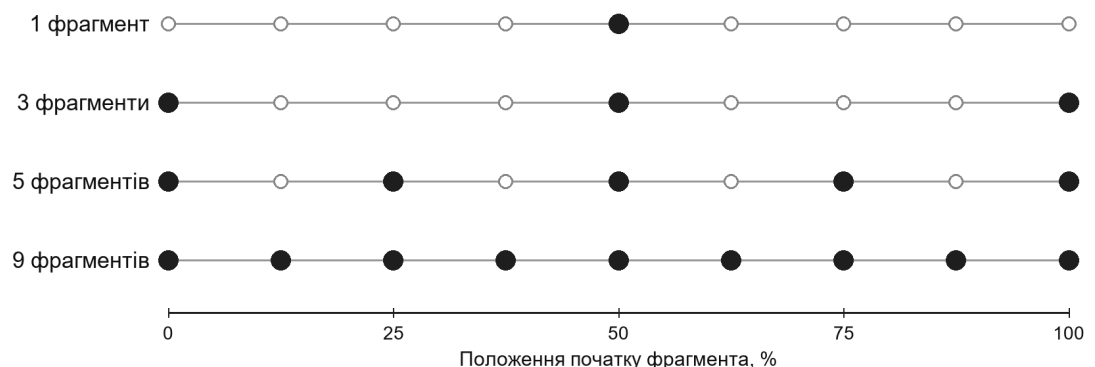


Рис. 1. Вкладені режими часового відбору. Чорні точки позначають використані позиції, порожні точки - спільну дев'ятиточкову сітку.

Основним, заздалегідь визначеним порівнянням була різниця ROC AUC між дев'ятьма фрагментами і одним центральним фрагментом на змішаній за якістю фінальній вибірці. Порівняння трьох і п'яти фрагментів з одним, а також дев'яти з п'ятьма вважали додатковими. Окремо аналізували вихідні відео без додаткового погіршення, помірно погіршені відео та невелику групу сильного погіршення. Остання містила лише 36 пар, тому її не використовували для основних висновків.

Гіпотеза дослідження полягала в тому, що рівномірне охоплення кількох ділянок підвищить ROC AUC відносно одного центрального фрагмента, але приріст матиме плато, оскільки сусідні фрагменти частково дублюють інформацію. Додаткова пояснювальна гіпотеза передбачала більшу мінливість прогнозів уздовж підроблених відео, ніж уздовж справжніх.

5.2. Набори даних і розподіл відео

Для навчання, внутрішньої перевірки та внутрішнього тесту використано відеодатасети FaceForensics++ c23 [11], Celeb-DF-v2 [12] і HiDF [13]. FaceForensics++ містить справжні записи та відео з кількома способами маніпуляції обличчям; версія c23 відображає стиснення, характерне для практичного поширення відео. Celeb-DF-v2 містить реальні та синтезовані відео публічних осіб з поліпшеними підробками. HiDF включає високоякісні матеріали, створені комерційними інструментами. Ці джерела відрізняються складом людей, генераторами, тривалістю, кодуванням і умовами знімання.

Навчальна частина містила 16 054 відео, валідаційна - 2 931, внутрішня тестова - 2 989. Розподіл виконувався не лише за файлами, а й за пов'язаними ідентичностями та вихідними записами, щоб похідні від одного матеріалу не опинилися одночасно у навчанні й тесті. Модель навчали тричі з різною випадковою ініціалізацією. Для кожного навчання зберігали стан з найкращим ROC AUC на внутрішній валідації, обчисленим за трьома фрагментами. У фінальному оцінюванні ймовірності трьох моделей усереднювали з однаковими вагами.

Зовнішнє оцінювання виконано з маніпуляціями DeeperForensics-1.0 [14]. Цей датасет створено для перевірки детекторів за реалістичних спотворень і містить вихідні підробки та версії з різним ступенем погіршення. Фінальна змішана вибірка складалася з 200 підроблених відео та 200 відповідних справжніх цільових відео FaceForensics++. Справжнє й підроблене відео в кожній парі відповідали тій самій цільовій особі й змісту. Ці файли та цільові ідентичності не перетиналися з частиною DeeperForensics, використаною раніше для вибору налаштувань.

Фінальна вибірка була незалежною за файлами та цільовими особами, але не повністю незалежною за походженням даних. Справжній клас представляли відкладені відео FaceForensics++, тоді як підроблений клас пройшов конвеєр DeeperForensics. Отже, клас частково пов'язаний із джерелом і постобробкою. Крім того, особи-донори могли повторюватися між допоміжною і фінальною частинами. Ці обмеження враховано в інтерпретації.

5.3. Архітектура незмінного детектора

Для всіх часових режимів застосовували одну й ту саму модель. На середньому кадрі кожного фрагмента детектор SCRFD знаходив найбільше обличчя [15]. Рамку збільшували у 1,5 рази відносно виявленої області, щоб зберегти найближче оточення і можливу межу вставлення. Та сама рамка використовувалася для всіх кадрів фрагмента, після чого зображення приводили до розміру 224 на 224 пікселі. На фінальних зовнішніх відео обличчя було знайдено в усіх перевірених фрагментах.

Кожний фрагмент містив 16 кадрів із кроком два кадри, тобто охоплював 31 позицію вихідної послідовності. Просторові ознаки кожного кадру виділяла ResNet-50 [16]. Паралельно для кожного кадру обчислювали логарифм модуля двовимірного швидкого перетворення Фур'є, який обробляла окрема згорткова гілка. Просторове та частотне подання об'єднували, додавали позиційне кодування та передавали до трьох шарів трансформерного кодера [17]. Після усереднення часових токенів лінійний класифікатор формував імовірність підробки.

Під час навчання до всіх кадрів одного фрагмента узгоджено застосовували випадкове зменшення й повторне збільшення роздільності, розмиття, зміну яскравості, контрасту, гамми та кількості кольорних рівнів. Таке погіршення не залежало від класу. Його завданням було зменшити залежність від конкретного кодування навчальних датасетів. У цій роботі ані архітектуру, ані кадрування, ані навчальні спотворення не змінювали; вони були фіксованою умовою для дослідження часового відбору.

5.4. Вкладені режими часового відбору

Схему вкладених режимів наведено на рис. 1. Дев'ять 16-кадрових фрагментів рівномірно розміщували між найпершою та найпізнішою допустимими позиціями. Режими з одним, трьома і п'ятьма фрагментами формували як вкладені підмножини цієї сітки. Якщо відео було коротшим за 31 кадр, 16 кадрів рівномірно відбирали з усієї його довжини. Така побудова забезпечила парне порівняння тих самих відео і дала змогу точно відновити всі режими з одного дев'ятифрагментного проходу.

5.5. Формування оцінки відео

Для кожного фрагмента спочатку усереднювали ймовірності, отримані трьома навченими моделями. Підсумковою оцінкою відео було арифметичне середнє оцінок усіх фрагментів вибраного режиму. Усі моделі та часові позиції мали однакову вагу.

Для дев'яти фрагментів додатково перевіряли медіану, максимум і середнє двох найбільших оцінок. Мінливість прогнозу уздовж відео описували стандартним відхиленням дев'яти фрагментних оцінок та різницею між найбільшою і найменшою з них.

5.6. Метрики та статистичний аналіз

Основною метрикою був ROC AUC, тобто ймовірність того, що випадково вибране підроблене відео отримає вищу оцінку, ніж випадково вибране справжнє [18]. ROC AUC не залежить від одного фіксованого порога, що важливо за перенесення моделі між датасетами. Збалансовану точність, чутливість і специфічність

обчислювали додатково, однак не використовували для вибору кількості фрагментів, оскільки поріг, перенесений з внутрішньої валідації, міг бути гірше відкалібрований на іншому джерелі.

Невизначеність різниці AUC оцінювали повторним формуванням 2 000 вибірок із поверненням [19]. Одиницею повторного відбору була ціла узгоджена пара зі справжнього та підробленого відео. Завдяки цьому обидва класи й зв'язок усередині пари зберігалися. Простими словами, метод багато разів моделює, який результат можна було б отримати за іншого набору таких самих пар.

Для кожного порівняння обчислювали двобічне р-значення. Воно відповідає на питання: якщо насправді різниці між режимами немає, наскільки ймовірно випадково отримати таку саму або ще більшу різницю? Для основного, заздалегідь визначеного порівняння дев'яти фрагментів з одним застосовували рівень значущості 0,05. Для додаткових порівнянь р-значення коригували методом Голма [20]. Цей метод послідовно посилює вимогу до значущості та обмежує ймовірність хоча б одного хибного висновку в групі додаткових перевірок.

Обчислювальний бюджет одного фрагмента приймали за 100%. За незмінної моделі та кадрів режимів з трьома, п'ятьма і дев'ятьма фрагментами потребували приблизно 300%, 500% і 900% цього бюджету відповідно.

5.7. Основне порівняння кількості фрагментів

Один центральний фрагмент забезпечив ROC AUC 0,8349 на фінальній змішаній вибірці. Додавання початку й кінця відео підвищило AUC до 0,8688. Для п'яти рівномірно розміщених фрагментів показник становив 0,8724, а для дев'яти - 0,8732.

Основний приріст дев'яти фрагментів відносно одного дорівнював 0,0383. Довірчий інтервал для цього приросту простягався від 0,0184 до 0,0592, а р-значення дорівнювало 0,001. Отже, поліпшення не пояснюється лише випадковим складом 200 узгоджених пар.

Три фрагменти дали приріст 0,0339 відносно одного та зберігали значущість після поправки Голма. П'ять фрагментів дали приріст 0,0375 і також зберігали значущість. Різниця між дев'ятьма та п'ятьма становила лише 0,0009 і не була підтверджена статистично. Це не доводить математичну рівноцінність двох режимів, оскільки окремий тест на не меншу ефективність не проводився, але показує відсутність виявленої переваги дев'яти фрагментів у цій вибірці.

Таблиця 1

Точність і відносний обчислювальний бюджет чотирьох режимів

Кількість фрагментів	ROC AUC	Зміна AUC проти 1	Частка приросту	Відносний бюджет, %
1	0,8349	0	0%	100%
3	0,8688	+0,0339	88,5%	300%
5	0,8724	+0,0375	97,7%	500%
9	0,8732	+0,0383	100%	900%

Три фрагменти забезпечили 88,5% приросту, який спостерігався між одним і дев'ятьма, за третини дев'ятифрагментного бюджету. П'ять фрагментів забезпечили 97,7% приросту за п'яти дев'ятих бюджету. Найбільший крок відбувся під час переходу від одного фрагмента до трьох; подальша крива швидко вийшла на плато.

5.8. Стійкість результату до погіршення якості

Патерн повторився у двох достатньо великих підгрупах. На вихідних відео перехід від одного до дев'яти фрагментів збільшив AUC на 0,0471. На помірно погіршених відео приріст становив 0,0344. В обох випадках додаткові порівняння з одним фрагментом залишалися значущими після поправки Голма, а підтвердженої переваги дев'яти над п'ятьма не було.

У групі сильного погіршення напрям ефекту був подібним, проте жодне порівняння не було підтверджено. Ця група містила лише 36 пар, тому широкий статистичний розкид не дає підстав стверджувати ані наявності, ані відсутності користі від додаткового покриття. Основні висновки спираються на змішану фінальну вибірку та дві повні 200-парні підгрупи.

На внутрішньому тесті AUC був високим уже для одного фрагмента, а абсолютний приріст від дев'яти становив близько 0,005. Отже, недостатнє часове покриття особливо проявилось після переходу до іншого джерела відео. Це узгоджується з тим, що на споріднених даних модель часто впевнено розпізнає більшість фрагментів, тоді як за зміни домену інформативність окремих ділянок стає нерівномірною.

5.9. Мінливість оцінок уздовж відео

Мінливість оцінок уздовж відео визначали за середнім стандартним відхиленням дев'яти фрагментних оцінок і середньою різницею між найбільшою та найменшою оцінками. Для справжніх відео ці показники становили 0,0806 і 0,2355, а для підроблених - 0,1510 і 0,4366 відповідно. Обидва показники були приблизно у 1,9 раза вищими для підробок, що свідчить про сильнішу часову неоднорідність їхніх ознак.

Аналіз окремих записів показав, що висока мінливість властива не кожному підробленому відео, але трапляється у них частіше й досягає більших значень. Тому середній прогноз за всім відео не можна відновити з

одного центрального фрагмента без ризику потрапити у нетипову ділянку. Рівномірне покриття зменшує цей ризик, оскільки поєднує фрагменти з різними позами, рухом і вираженістю артефактів.

Цей результат пояснює, чому найбільший приріст виник уже після додавання початку й кінця. Три точки різко збільшують часову відстань між спостереженнями. Додаткові проміжні точки уточнюють середнє, але частково дублюють інформацію, тому приріст після п'яти фрагментів майже зникає.

5.10. Правила об'єднання дев'яти оцінок

На фінальній змішаній вибірці арифметичне середнє забезпечило AUC 0,8732. Для медіани AUC становив 0,8707, для середнього двох найбільших оцінок - 0,8715, для максимуму - 0,8677. Жодне альтернативне правило не перевершило середнє після поправки на множинні порівняння.

Середнє двох найбільших оцінок підвищувало чутливість до підробок, але водночас збільшувало частку хибних спрацювань на справжніх відео. Максимум посилював цю залежність від одиничного фрагмента ще більше. Медіана була стійкішою до крайніх значень, проте могла відкидати коротку інформативну ділянку. За відсутності підтвердженої переваги складнішого правила звичайне середнє залишено основним.

5.11. Інтерпретація основного результату

Основний результат полягає не просто в тому, що більше кадрів дають вищий AUC. Експеримент визначив форму залежності для незмінної моделі: різкий приріст від одного до трьох, невелике поліпшення до п'яти та відсутність виявленої користі від подальшого збільшення до дев'яти. Вкладена сітка дає змогу пов'язати цю залежність саме з часовим покриттям, а не з повторним випадковим вибором або іншою архітектурою.

Один центральний фрагмент є привабливим через мінімальну вартість, проте не має спеціальних підстав бути найінформативнішою частиною відео. У середині запису обличчя може бути нерухомим, відвернутим, перекритим або зображеним у момент, коли артефакт заміни слабкий. Початок і кінець додають віддалені в часі стани. Саме тому три точки забезпечили основну частину приросту.

Більша внутрішньовідео мінливість у підробках дає механістичне пояснення. Маніпуляція може бути стабільною на одних позах і менш стабільною на інших; межа вставлення, деталізація шкіри та узгодженість руху змінюються разом із положенням обличчя. Модель реагує на ці зміни, тому її ймовірності коливаються сильніше. Усереднення віддалених фрагментів зменшує вплив як випадково легкого, так і випадково складного відрізка.

Водночас часовий відбір не усуває міждасетний зсув. Навіть за дев'яти фрагментів зовнішній AUC залишався нижчим за внутрішній. Отже, ширше покриття виправляє лише одну складову помилки - нерепрезентативність короткого відрізка. Відмінності генераторів, постобробки, людей і джерел відео потребують окремих методів узагальнення.

5.12. Співвідношення з попередніми роботами

Рекурентні, згорткові та трансформерні методи довели, що часові залежності містять корисні ознаки підробки [3-9]. Поточне дослідження розглядає інший рівень часової структури. Кожен окремий 16-кадровий фрагмент уже аналізується трансформером, але додатково перевірено, скільки таких коротких спостережень потрібно розподілити вздовж цілого відео.

Рівномірний поділ нагадує принцип Temporal Segment Networks [10], однак тут він використаний не для навчання розпізнавача дій, а для контрольованого оцінювання дипфейків після завершення навчання. Наукове значення полягає у парному вимірюванні точності та вартості, у визначенні плато і в пов'язанні приросту з реальною мінливістю фрагментних прогнозів. Таким чином, робота доповнює дослідження часових архітектур аналізом протоколу їх застосування до цілого відео.

Результат TALL показує, що для відеодетекції важливий компроміс між збереженням часових зв'язків і вартістю [9]. Отримане плато уточнює цей компроміс для іншої моделі: збільшувати часовий бюджет до максимуму не завжди доцільно. Водночас висновок не означає, що п'ять фрагментів є універсальним числом для всіх детекторів. Воно залежить від тривалості відео, довжини одного фрагмента, архітектури та характеру маніпуляцій.

5.13. Практичне використання

За жорсткого обмеження часу доцільно використовувати три фрагменти: початок, середину й кінець. Цей режим отримав більшу частину максимально спостережуваного приросту і потребує лише третини обчислень дев'ятифрагментного режиму. Для пакетної експертизи, де помірне збільшення часу прийнятне, п'ять фрагментів є обережнішим вибором. Воно майже досягло дев'ятифрагментного результату без чотирьох додаткових проходів.

Дев'ять фрагментів можуть залишатися виправданими, якщо обчислення виконують паралельно, якщо потрібні щільніші пояснення вздовж відео або якщо запис довший і складніший за досліджені. Проте для коротких еталонних відео поточні дані не підтвердили його перевагу над п'ятьма. У робочій системі кількість фрагментів варто визначати разом із допустимою затримкою, ціною хибного пропуску та типом відеопотоку.

Окремий напрям становить адаптивний відбір. Замість фіксованої кількості система може спочатку обробити три фрагменти, оцінити їхню узгодженість і додавати нові лише за великого розкиду або близькості рішення до порога. Така стратегія безпосередньо впливає з виявленої часової неоднорідності, але в цій роботі її не перевіряли, тому вона є напрямом подальшого дослідження, а не отриманим результатом.

5.14. Обмеження дослідження

Отримані результати стосуються відео тривалістю приблизно від 8,5 до 49 секунд, однієї архітектури та фіксованих 16-кадрових фрагментів. Тому виявлене плато біля п'яти фрагментів не слід без окремої перевірки переносити на інші моделі або багатовихідні записи. Для довгих реальних відео один фрагмент не може представляти весь запис; подальше дослідження має враховувати тривалість, зміни сцен та невизначеність прогнозу. Крім того, група з найсильнішим погіршенням якості містила лише 36 пар, тому її результати мають описовий характер.

6. Висновки та перспективи подальших досліджень

Для незмінного детектора один центральний фрагмент гірше представляв коротке зовнішнє відео, ніж рівномірне покриття кількох часових ділянок. ROC AUC зріс з 0,8349 для одного фрагмента до 0,8688 для трьох, 0,8724 для п'яти та 0,8732 для дев'яти. Основний приріст дев'яти проти одного становив 0,0383 і був статистично підтверджений.

Залежність мала виражене плато. Три фрагменти отримали 88,5% максимального спостережуваного приросту за третини дев'ятифрагментного бюджету. П'ять отримали 97,7% приросту, а переваги дев'яти над п'ятьма не виявлено. Тому три фрагменти доцільні для економного режиму, а п'ять - як обережна конфігурація поблизу насичення.

Підроблені відео мали приблизно у 1,9 раза більшу мінливість оцінок уздовж часу. Це пояснює ризик одного фрагмента: він може потрапити у ділянку зі слабкими або нетиповими ознаками. Звичайне середнє дев'яти оцінок не поступилося медіані, максимуму та середньому двох найбільших значень.

Виявлене плато стосується коротких еталонних відео й однієї архітектури. Наступним кроком має бути адаптивний протокол для багатовихідних реальних записів, у якому кількість і положення фрагментів залежать від тривалості, змін сцен, якості обличчя та незгоди між початковими прогнозами.

Внесок авторів

Усі автори брали участь у концептуалізації дослідження, аналізі результатів, підготовці та редагуванні рукопису і схвалили його остаточну версію.

Декларація про штучний інтелект

Генеративну систему штучного інтелекту використано для мовно-стилістичного редагування та перевірки оформлення рукопису. Її не використовували для отримання експериментальних результатів, побудови наборів даних, статистичного аналізу або формування наукових висновків. Авторі перевірили остаточний текст і несуть відповідальність за його зміст.

Конфлікт інтересів

Авторі заявляють про відсутність конфлікту інтересів і підтверджують, що під час підготовки цієї роботи не існувало комерційних, фінансових або інших взаємовідносин, здатних вплинути на результати дослідження чи їх інтерпретацію.

Список використаної літератури

1. Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*. 2020. Vol. 64. P. 131-148. DOI: 10.1016/j.inffus.2020.06.014.
2. Yan Z., Zhang Y., Yuan X., Lyu S., Wu B. DeepfakeBench: A comprehensive benchmark of deepfake detection. *Advances in Neural Information Processing Systems*. 2023. Vol. 36. Datasets and Benchmarks Track. DOI: 10.52202/075280-0201.
3. Guera D., Delp E. J. Deepfake video detection using recurrent neural networks. *15th IEEE International Conference on Advanced Video and Signal Based Surveillance*. 2018. P. 1-6. DOI: 10.1109/AVSS.2018.8639163.
4. Sabir E., Cheng J., Jaiswal A., AbdAlmageed W., Masi I., Natarajan P. Recurrent convolutional strategies for face manipulation detection in videos. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019. P. 80-87.
5. Haliassos A., Vougioukas K., Petridis S., Pantic M. Lips don't lie: A generalisable and robust approach to face forgery detection. *Proceedings of CVPR*. 2021. P. 5039-5049. DOI: 10.1109/CVPR46437.2021.00500.
6. Zheng Y., Bao J., Chen D., Zeng M., Wen F. Exploring temporal coherence for more general video face forgery detection. *Proceedings of ICCV*. 2021. P. 15044-15054.
7. Haliassos A., Mira R., Petridis S., Pantic M. Leveraging real talking faces via self-supervision for robust forgery detection. *Proceedings of CVPR*. 2022. P. 14950-14962.
8. Wang Z., Bao J., Zhou W., Wang W., Li H. AltFreezing for more general video face forgery detection. *Proceedings of CVPR*. 2023. P. 4129-4138.
9. Xu Y., Liang J., Jia G., Yang Z., Zhang Y., He R. TALL: Thumbnail layout for deepfake video detection. *Proceedings of ICCV*. 2023. P. 22658-22668.
10. Wang L., Xiong Y., Wang Z., Qiao Y., Lin D., Tang X., Van Gool L. Temporal segment networks: Towards good practices for deep action recognition. *Computer Vision - ECCV 2016. Lecture Notes in Computer Science*. 2016. Vol. 9912. P. 20-36. DOI: 10.1007/978-3-319-46484-8_2.
11. Roessler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Niessner M. FaceForensics++: Learning to detect manipulated facial images. *Proceedings of ICCV*. 2019. P. 1-11. DOI: 10.1109/ICCV.2019.00009.
12. Li Y., Yang X., Sun P., Qi H., Lyu S. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. *Proceedings of CVPR*. 2020. P. 3207-3216.
13. Kang C., Jeong S., Lee J., Choi D., Woo S. S., Han J. HiDF: A human-indistinguishable deepfake dataset. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2025. Vol. 2. P. 5527-5538. DOI: 10.1145/3711896.3737399.
14. Jiang L., Li R., Wu W., Qian C., Loy C. C. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. *Proceedings of CVPR*. 2020. P. 2889-2898. DOI: 10.1109/CVPR42600.2020.00296.

15. Guo J., Deng J., Lattas A., Zafeiriou S. Sample and computation redistribution for efficient face detection. International Conference on Learning Representations. 2022. URL: <https://openreview.net/forum?id=RhB1AdoFfGE>.
16. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. Proceedings of CVPR. 2016. P. 770-778. DOI: 10.1109/CVPR.2016.90.
17. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998-6008.
18. Hanley J. A., McNeil B. J. The meaning and use of the area under a receiver operating characteristic curve. Radiology. 1982. Vol. 143, No. 1. P. 29-36. DOI: 10.1148/radiology.143.1.7063747.
19. Efron B., Tibshirani R. J. An introduction to the bootstrap. New York: Chapman and Hall/CRC, 1993. 456 p. DOI: 10.1201/9780429246593.
20. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979. Vol. 6, No. 2. P. 65-70. DOI: 10.2307/4615733.

References

1. Tolosana R., Vera-Rodriguez R., Fierrez J., Morales A., Ortega-Garcia J. Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion. 2020. Vol. 64. P. 131-148. DOI: 10.1016/j.inffus.2020.06.014.
2. Yan Z., Zhang Y., Yuan X., Lyu S., Wu B. DeepfakeBench: A comprehensive benchmark of deepfake detection. Advances in Neural Information Processing Systems. 2023. Vol. 36. Datasets and Benchmarks Track. DOI: 10.52202/075280-0201.
3. Guera D., Delp E. J. Deepfake video detection using recurrent neural networks. 15th IEEE International Conference on Advanced Video and Signal Based Surveillance. 2018. P. 1-6. DOI: 10.1109/AVSS.2018.8639163.
4. Sabir E., Cheng J., Jaiswal A., AbdAlmageed W., Masi I., Natarajan P. Recurrent convolutional strategies for face manipulation detection in videos. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019. P. 80-87.
5. Haliassos A., Vougioukas K., Petridis S., Pantic M. Lips don't lie: A generalisable and robust approach to face forgery detection. Proceedings of CVPR. 2021. P. 5039-5049. DOI: 10.1109/CVPR46437.2021.00500.
6. Zheng Y., Bao J., Chen D., Zeng M., Wen F. Exploring temporal coherence for more general video face forgery detection. Proceedings of ICCV. 2021. P. 15044-15054.
7. Haliassos A., Mira R., Petridis S., Pantic M. Leveraging real talking faces via self-supervision for robust forgery detection. Proceedings of CVPR. 2022. P. 14950-14962.
8. Wang Z., Bao J., Zhou W., Wang W., Li H. AltFreezing for more general video face forgery detection. Proceedings of CVPR. 2023. P. 4129-4138.
9. Xu Y., Liang J., Jia G., Yang Z., Zhang Y., He R. TALL: Thumbnail layout for deepfake video detection. Proceedings of ICCV. 2023. P. 22658-22668.
10. Wang L., Xiong Y., Wang Z., Qiao Y., Lin D., Tang X., Van Gool L. Temporal segment networks: Towards good practices for deep action recognition. Computer Vision - ECCV 2016. Lecture Notes in Computer Science. 2016. Vol. 9912. P. 20-36. DOI: 10.1007/978-3-319-46484-8_2.
11. Roessler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Niessner M. FaceForensics++: Learning to detect manipulated facial images. Proceedings of ICCV. 2019. P. 1-11. DOI: 10.1109/ICCV.2019.00009.
12. Li Y., Yang X., Sun P., Qi H., Lyu S. Celeb-DF: A large-scale challenging dataset for DeepFake forensics. Proceedings of CVPR. 2020. P. 3207-3216.
13. Kang C., Jeong S., Lee J., Choi D., Woo S. S., Han J. HiDF: A human-indistinguishable deepfake dataset. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2025. Vol. 2. P. 5527-5538. DOI: 10.1145/3711896.3737399.
14. Jiang L., Li R., Wu W., Qian C., Loy C. C. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. Proceedings of CVPR. 2020. P. 2889-2898. DOI: 10.1109/CVPR42600.2020.00296.
15. Guo J., Deng J., Lattas A., Zafeiriou S. Sample and computation redistribution for efficient face detection. International Conference on Learning Representations. 2022. URL: <https://openreview.net/forum?id=RhB1AdoFfGE>.
16. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. Proceedings of CVPR. 2016. P. 770-778. DOI: 10.1109/CVPR.2016.90.
17. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems. 2017. Vol. 30. P. 5998-6008.
18. Hanley J. A., McNeil B. J. The meaning and use of the area under a receiver operating characteristic curve. Radiology. 1982. Vol. 143, No. 1. P. 29-36. DOI: 10.1148/radiology.143.1.7063747.
19. Efron B., Tibshirani R. J. An introduction to the bootstrap. New York: Chapman and Hall/CRC, 1993. 456 p. DOI: 10.1201/9780429246593.
20. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 1979. Vol. 6, No. 2. P. 65-70. DOI: 10.2307/4615733.

Надійшла до редакції: 05.03.2026

Прийнята до друку: 15.09.2026

Опубліковано: 30.09.2026